CS5200: DATABASE MANAGEMENT SYSTEMS
PROF. KATHLEEN DURANT

Perturbation Compendium for Biological and Pharmaceutical Research

Students: Li McCarthy, Pengcheng Wu
Blackboard Group Name: McCarthyWu

April 17, 2019

Contents

1 INTRODUCTION

1.1 TOP-LEVEL DESCRIPTION

In biology, a perturbation is any change in gene expression which can result from genetic disturbance, small molecule activity, or disease. The CMap, or âĂIJConnectivity MapâĂİ (https://www.broadinstitute.org/connectivity-map-cmap), is a large-scale dataset maintained by the Broad institute which documents and cross-compares these perturbations. It facilitates the discovery of connections between genes,drugs, and diseases. CLUE ConnectivityMap (https://clue.io/), is a integrated database environment also maintained by the institute, allowing users to query differences in gene transcription, which is based on 9 core cell lines across around 30 tissues. Computational biologists can access data using data APIs at https://clue.io/api. The perturbations include small molecules and other genetic perturbagen like shRNAs, cDNAs and biologics. They are treated on target genes for monitoring the downstream consequences of gene expression.

This project represents a small proof of concept database which integrates data from the CMap datasets with other publically available biotechnology data in able to link the following data domains: genes, diseases, cell lines, tissues, publications, and perturbations.

1.2 FUNCTIONALITY TO THE USER

This database is designed to provide a tool for a researcher to add biological data to a database, and find links among that data which may be useful for drug development or other correlational research. There should be two types of end user: 1) an administrator who can create/read/update/delete tuple data, and 2) a user who can only read. Presumably, there will be an external system for submitting documents to review and insert. The user should be able to search by specific metrics, for example, gene, cell line, disease, and literature.

2 README

2.1 Libraries, Software, Technical Specifications

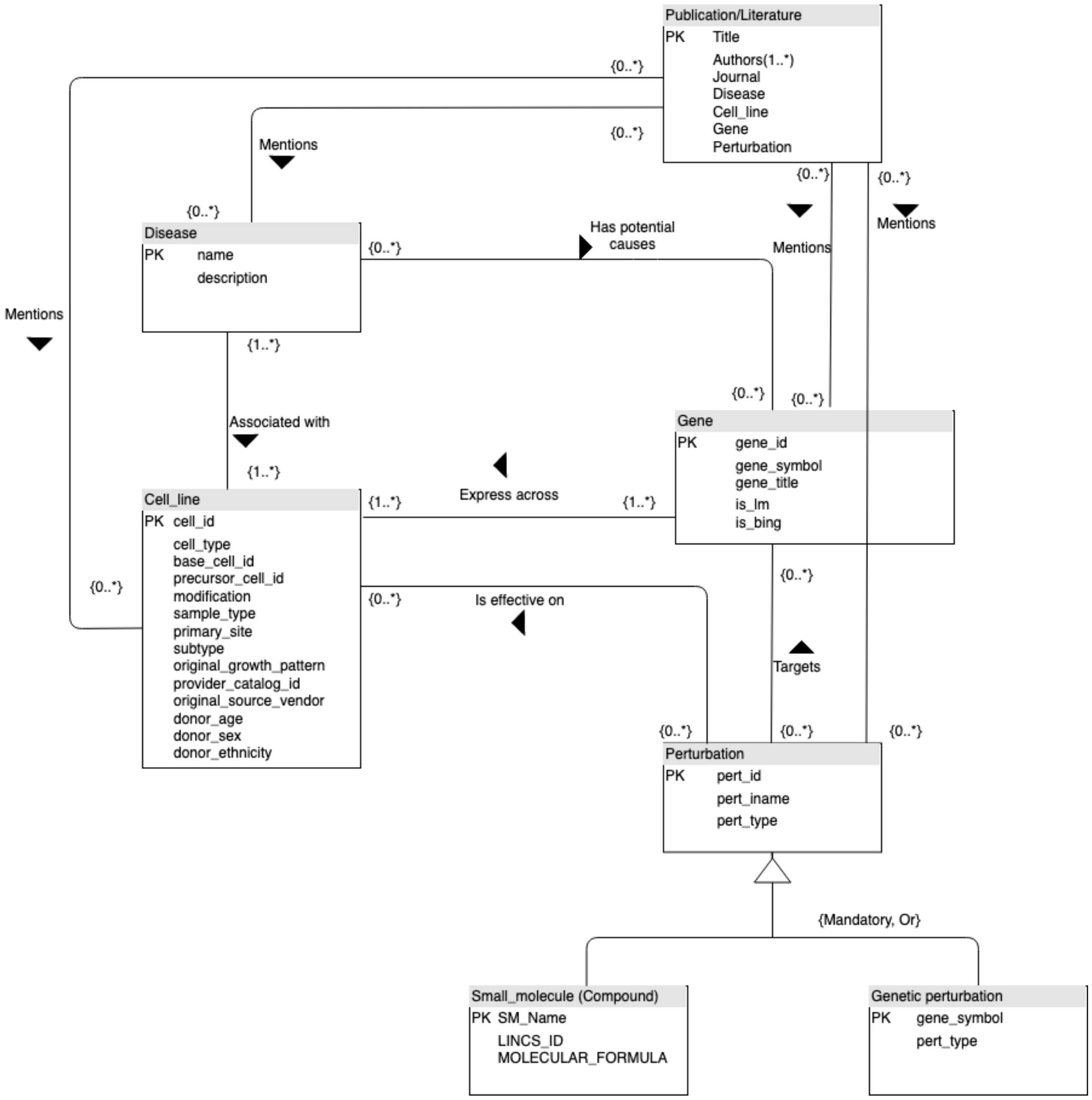
The back-end for this project is developed in MySQL 8.0.13, and the front end with Python 3.6.8. For the purpose of the project, the program runs from the command line to avoid requiring a specific OS.

Libraries used:
· mySQLdb (https://mysqlclient.readthedocs.io/user_guide.html)
· PyQT5 (https://pypi.org/project/PyQt5/)
· matplotlib.PyPlot (https://pypi.org/project/matplotlib/)

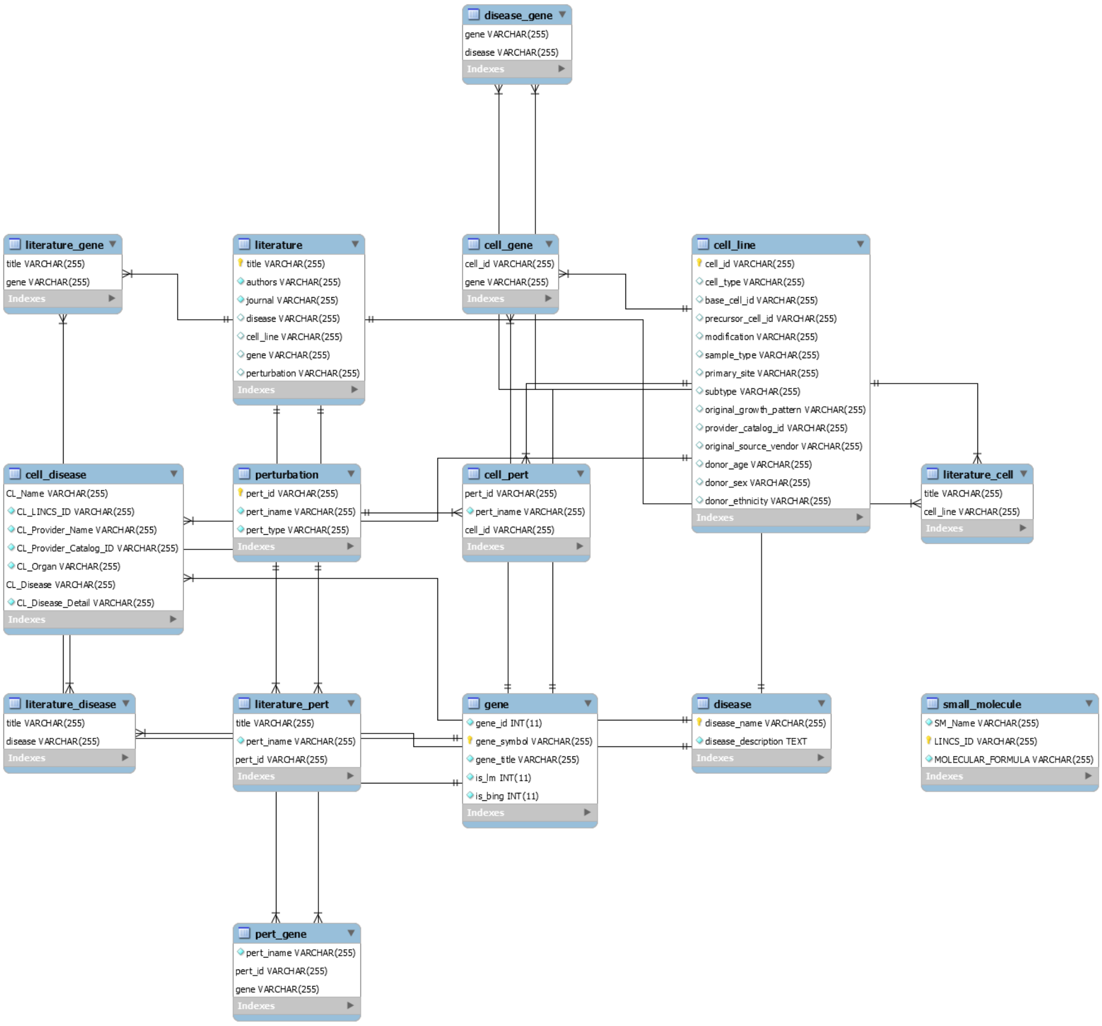
2.2 Usage

User can login as administrator with credentials username: "admin", password: "password". Please set up a user on a sql server running on localhost with those credentials, then run the included schema dump. After that, run the GUI with <python bio_compendium.py>. The database intentionally does not store or protect login information so as to narrow the scope of the project. Find more details in the "User Workflow" section.

3 Conceptual Database Design



4 Logical Database Design



5 User Workflow

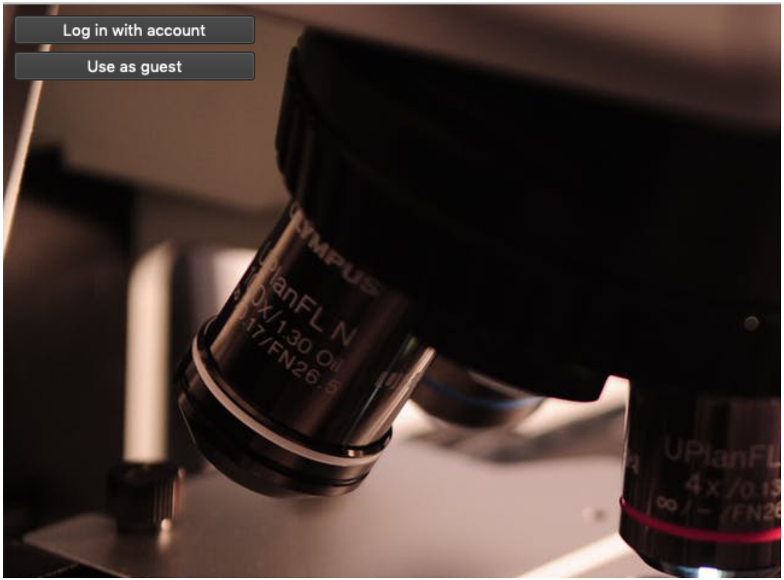
5.1 Setup

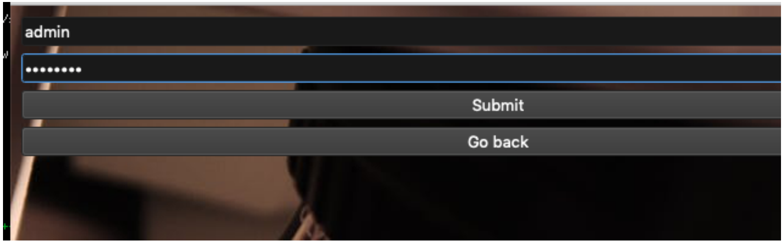
For technology, please refer to the required packages in section 2.
Please ensure that a mySQL server is running on localhost with a user with read/write access and credentials ("admin"/"password").

Run the schema named <cmap_project2.0.sql> to create the database. The GUI can be run from the command line with <python bio_compendium.py>.

5.2 Login

The program allows read-only access for a guest user and read-write access for a logged-in user. Again, the username and password are hardcoded.

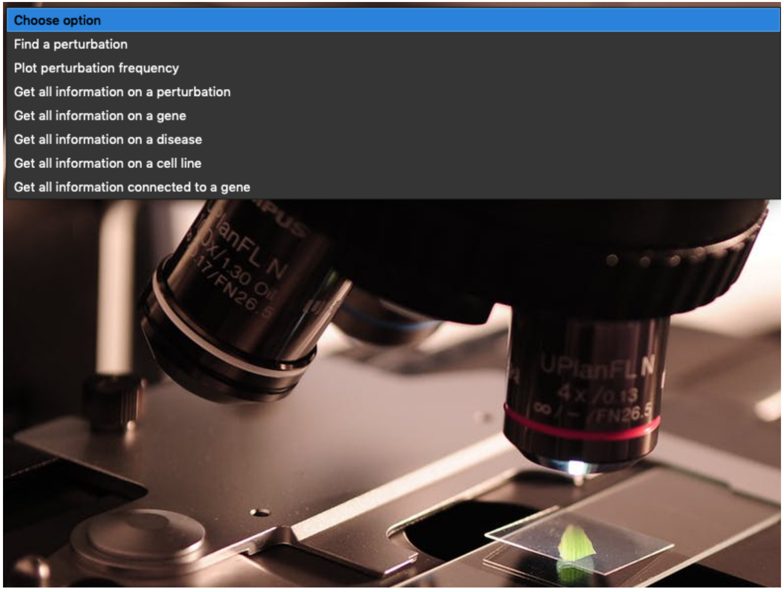



admin
••••••••

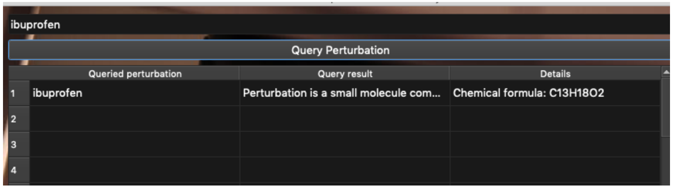
To login with write access, use username "admin" and password "password."

5.3 Read-only

As a guest, one can do the following:

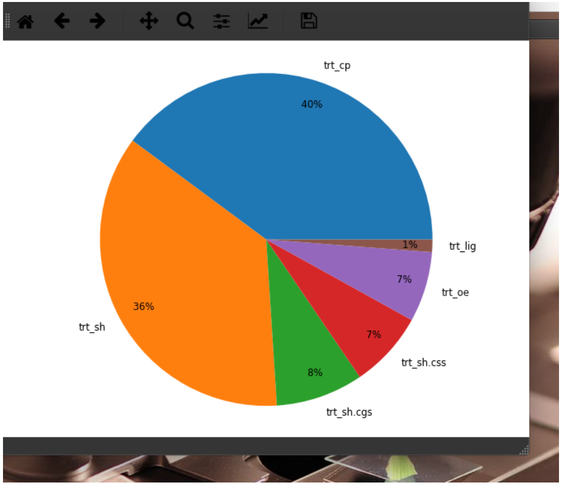

Choose option
Find a perturbation
Plot perturbation frequency
Get all information on a perturbation
Get all information on a gene
Get all information on a disease
Get all information on a cell line
Get all information connected to a gene

· **Find a perturbation**: Given a perturbation name, this will query the database and return whether it is present or not. If there is additional information about the perturbation in the small molecule table, it will retrieve the chemical formula of the perturbation. If not, it will specify that the perturbation is present in the database but information cannot be found.

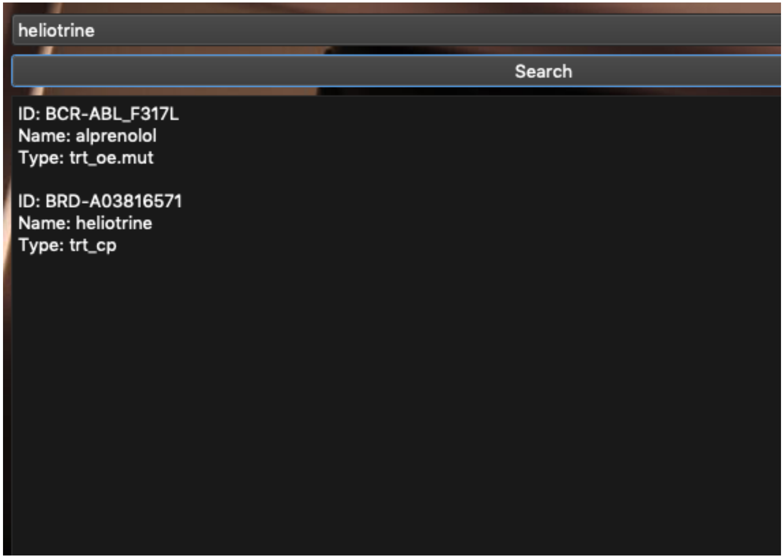


	Queried perturbation	Query result	Details
1	ibuprofen	Perturbation is a small molecule com...	Chemical formula: C13H18O2
2			
3			
4			

· **Plot perturbation frequency**: This shows a plot of the most frequent perturbations in the database.

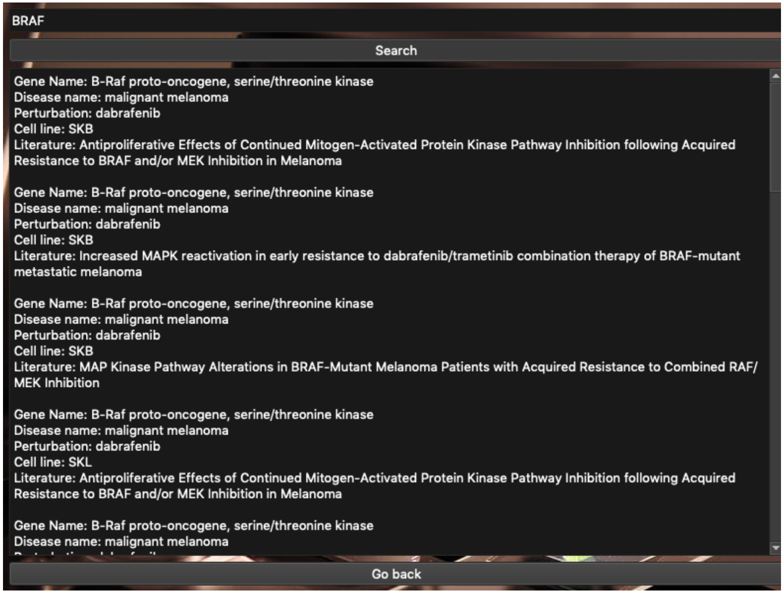


· **Get all information on a perturbation/gene/cell line**: These operations do a SQL query for all of the fields matching corresponding to the record requested. The perturbation, gene, or cell line names are automatically populated when this option is selected.



```
ID: BCR-ABL_F317L
Name: alprenolol
Type: trt_oe.mut

ID: BRD-A03816571
Name: heliotrine
Type: trt_cp
```

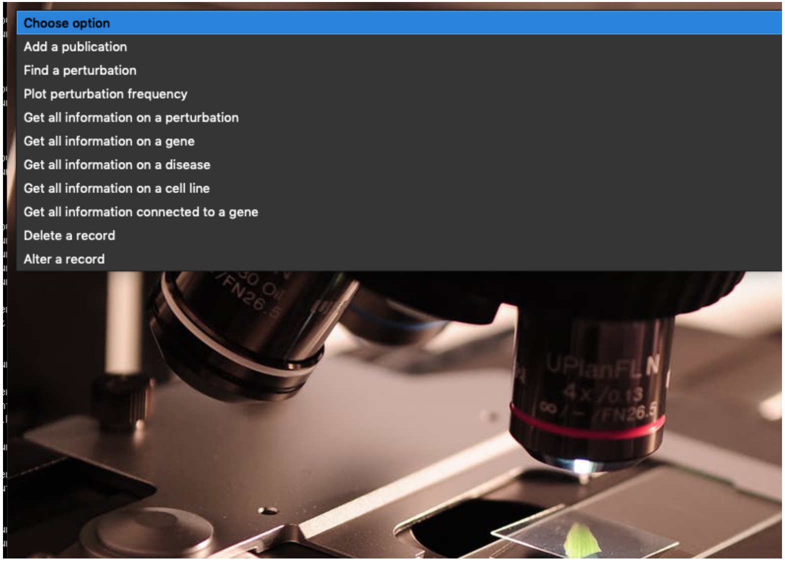
· **Get all information related to a gene**: Aggregates information from disease, perturbation, cell line, and literature.



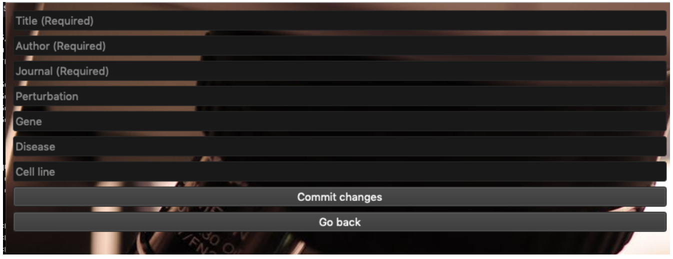
```
Gene Name: B-Raf proto-oncogene, serine/threonine kinase
Disease name: malignant melanoma
Perturbation: dabrafenib
Cell line: SKB
Literature: Antiproliferative Effects of Continued Mitogen-Activated Protein Kinase Pathway Inhibition following Acquired
Resistance to BRAF and/or MEK Inhibition in Melanoma

Gene Name: B-Raf proto-oncogene, serine/threonine kinase
Disease name: malignant melanoma
Perturbation: dabrafenib
Cell line: SKB
Literature: Increased MAPK reactivation in early resistance to dabrafenib/trametinib combination therapy of BRAF-mutant
metastatic melanoma

Gene Name: B-Raf proto-oncogene, serine/threonine kinase
Disease name: malignant melanoma
Perturbation: dabrafenib
Cell line: SKB
Literature: MAP Kinase Pathway Alterations in BRAF-Mutant Melanoma Patients with Acquired Resistance to Combined RAF/
MEK Inhibition

Gene Name: B-Raf proto-oncogene, serine/threonine kinase
Disease name: malignant melanoma
Perturbation: dabrafenib
Cell line: SKL
Literature: Antiproliferative Effects of Continued Mitogen-Activated Protein Kinase Pathway Inhibition following Acquired
Resistance to BRAF and/or MEK Inhibition in Melanoma

Gene Name: B-Raf proto-oncogene, serine/threonine kinase
Disease name: malignant melanoma
```

5.4 Read-Write

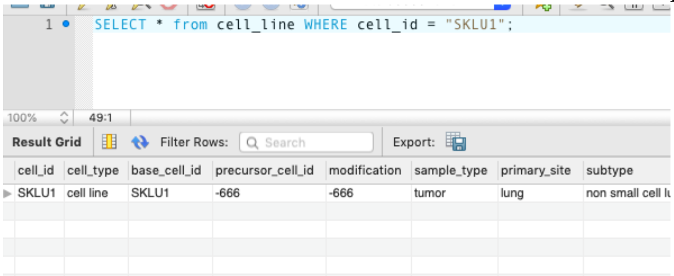
Read-write functionality adds the following options:


Choose option
Add a publication
Find a perturbation
Plot perturbation frequency
Get all information on a perturbation
Get all information on a gene
Get all information on a disease
Get all information on a cell line
Get all information connected to a gene
Delete a record
Alter a record

· **Add a publication**: This allows a user to insert a new piece of literature into the database with up to one linked perturbation, gene, and cell line.

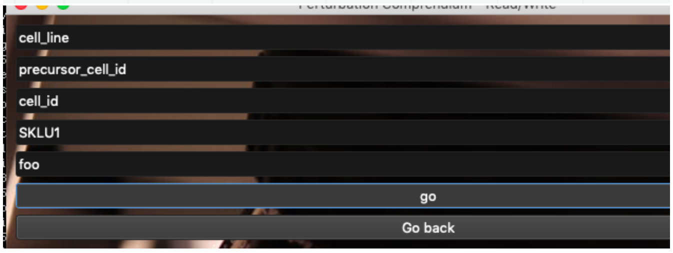


· **Alter a record**: This allows the administrator to alter a pre-existing record in the database.



```
SELECT * from cell_line WHERE cell_id = "SKLU1";
```

cell_id	cell_type	base_cell_id	precursor_cell_id	modification	sample_type	primary_site	subtype
SKLU1	cell line	SKLU1	-666	-666	tumor	lung	non small cell l



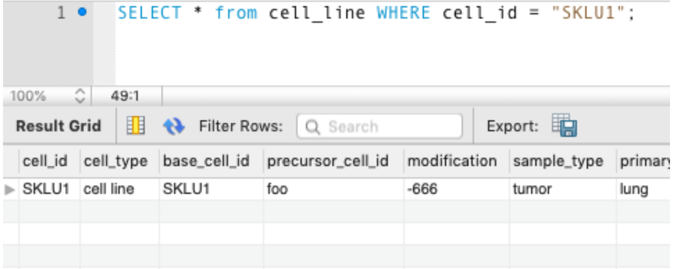
cell_line

precursor_cell_id

cell_id

SKLU1

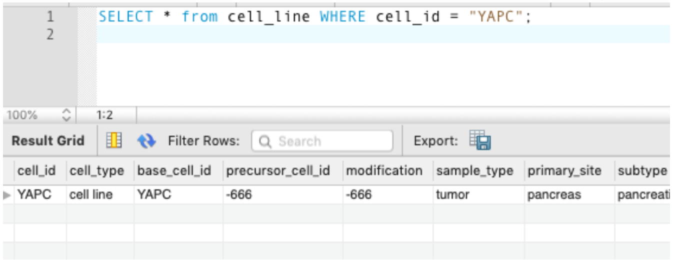
foo



```
SELECT * from cell_line WHERE cell_id = "SKLU1";
```

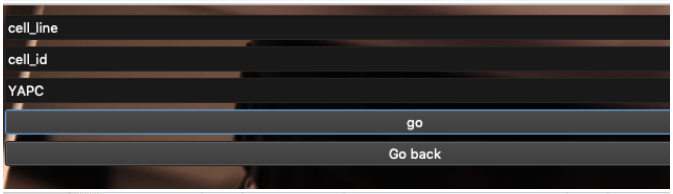
cell_id	cell_type	base_cell_id	precursor_cell_id	modification	sample_type	primary
SKLU1	cell line	SKLU1	foo	-666	tumor	lung

· **Delete a record**: This allows the administrator to delete a record currently in the database. Deletions are not allowed if the records are already cross-referenced in other tables.



```
SELECT * from cell_line WHERE cell_id = "YAPC";
```

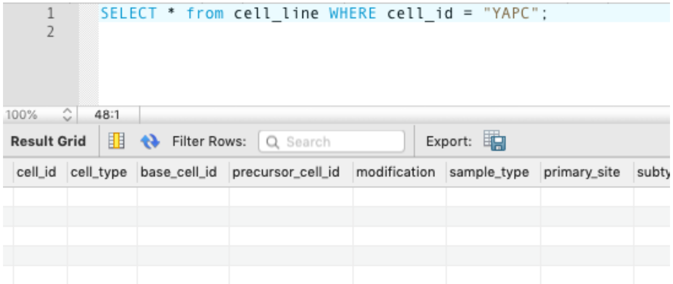
cell_id	cell_type	base_cell_id	precursor_cell_id	modification	sample_type	primary_site	subtype
YAPC	cell line	YAPC	-666	-666	tumor	pancreas	pancreati



cell_line

cell_id

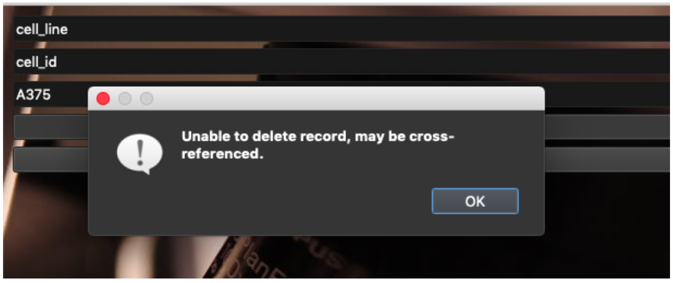
YAPC



```
SELECT * from cell_line WHERE cell_id = "YAPC";
```

cell_id	cell_type	base_cell_id	precursor_cell_id	modification	sample_type	primary_site	subty

If the record is cross-referenced, the user cannot delete it.



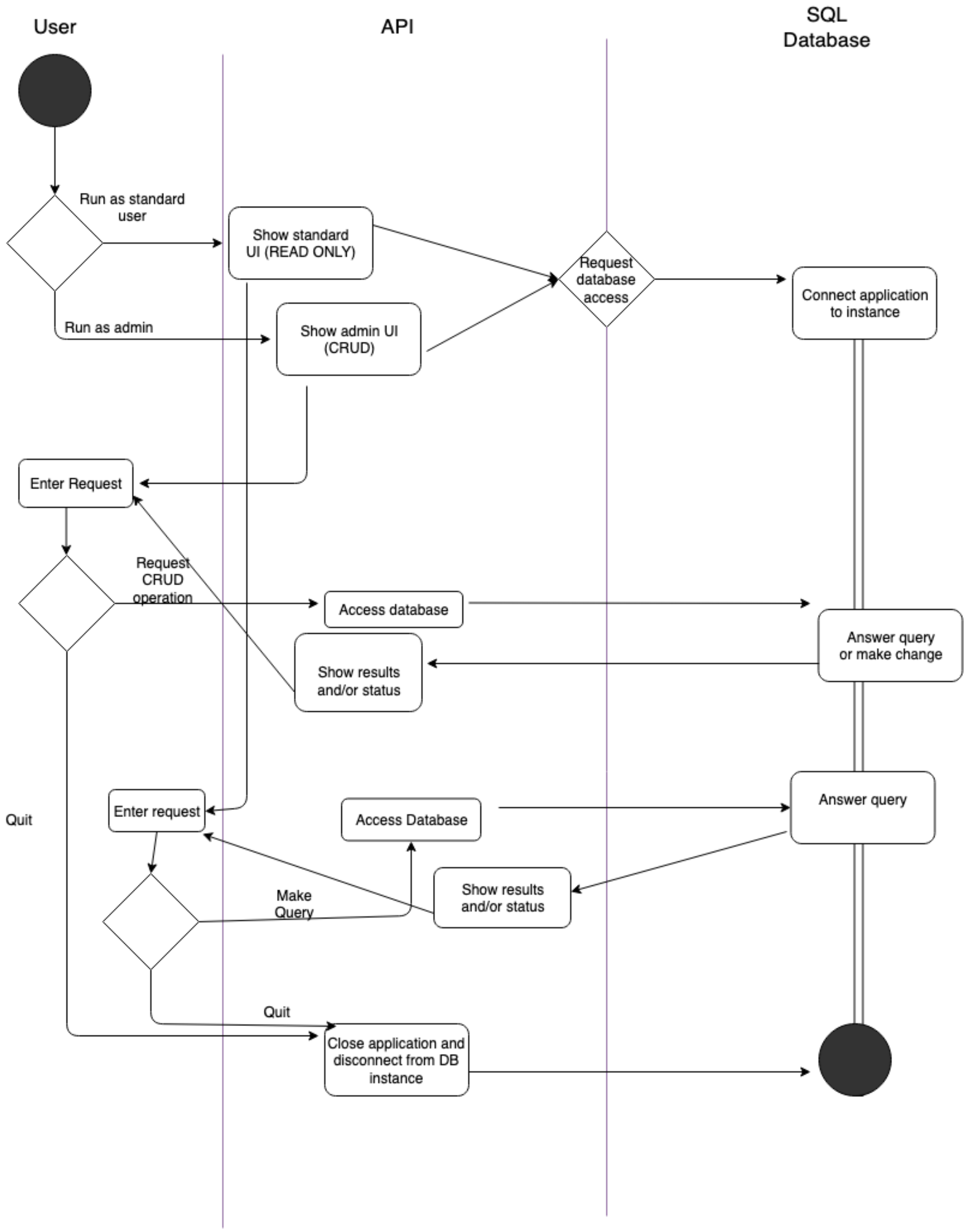
cell_line

cell_id

A375

Unable to delete record, may be cross-referenced.

6 Activity Diagram



7 Lessons Learned

7.1 Technical Expertise Gained

This project contained many firsts for both group members. Including the class assignment, it was our second time connecting a database to a front-end program, and our first time writing a GUI in python.

7.2 Group work insights, time management insights, data domain insights

7.2.1 Group work

Group work went fairly smoothly. We wrote the project proposal and planned out the conceptual design at the same time. Pengcheng did the majority of dataset pre-processing, while Li designed the user workflow. After that, Pengcheng designed and implemented the majority of the back-end code and Li designed and implemented the majority of the front-end code. Concurrency control was managed through git with minimal issues.

The biggest issue encountered involved concurrency and file types on our different machines– we encountered some issues with data import (for example, source folders on Windows vs OSX, different types of newlines and carriage returns interfering with data reading, etc). These could definitely have been solved by sitting down and making sure all of the data was imported properly first, then putting a moratorium on adding new records through large-scale import as opposed to insertion with SQL queries. Another small challenge is that the members collected data from multiple data resources rather than downloading a data package from online data community, so there are some extra efforts to create connection between data gained with different approaches.

7.2.2 Time Management

As mentioned previously, the entire project work started with team members' first meeting to determine the topic and work schedule. Then we began to design the proposal of the project. After getting the feedback of the proposal, we were encouraged by the professor and became more strong willed to insist on doing what we expected to achieve. During the semester, both members constantly updated their understanding about database design and the project background. At the same time, the members were also doing attempts to collect data from different resources and standardized them into a unified format so that some possible reference relationship can be defined later. By the first week of presentation, the team has finished majority part of database setup and about three functions consisting inserting data, extracting data and judging the perturbation type of a given perturbation and show related details if it satisfies the requirement of showing. By the second week of presentation, the project work was almost done with about ten operations covered insert, read, update, and delete queries (CRUD). Visualization function was also available when analyzing the perturbation category. The only work left by the last week was the final report.

7.2.3 Data Domain

The project mainly has five data domains including literature, gene, disease, cell line, and perturbation. The perturbation domain can be divided into small molecule and genetic perturbation according to the category. Every literature should provide its title, authors and journal name as basic information and some keywords among the other four data domains as optional information. Gene is taken from human gene database and covers almost all common genes related with recorded human diseases. It consist both gene symbol and gene title for study. Disease mainly covers the most common series disease which has a reasonable interpretation in gene level, like cancer or other congenital diseases. Perturbation means a certain kind of alteration in the function of a biological system. Specifically speaking, a genetic perturbation represent alteration in function due to change in gene expression in related pathways, while small molecule represent chemical compound served as non-genetic treatment such as pharmaceuticals. A cell line is a permanently established cell culture that will proliferate indefinitely given appropriate fresh medium and space. Cell lines are an essential part of in-vitro assays, and tend to have large bodies of research already present on what in-vivo systems they are similar to. In our database system, it will provide a cell line's basic information and the provider's public data.

7.3 Realized or contemplated alternative design and approaches

We created a lot of extra work for ourselves by using a MySQL database rather than a NoSQL one. Almost every relationship between tables was many-many simply due to the many cross-references required. However, we chose to focus on refining our skill with MySQL.

Additionally, we picked quite a broad topic through the inclusion of user/admin privileges and literature. While we believe that the inclusion of both of these aspects helped to demonstrate our program as a fully-realized project, we had to make a tradeoff between including this extra functionality and making more interesting operations on the relationship between disease, cell line, gene, and perturbation. Given the opportunity to revisit this project, I believe that we would lean further into the data visualization angle (for example, expressing the connectivity between perturbations, genes, and disease, or calculating more frequency metrics).

7.4 Document any code not working in this section

None!

8 Future Work

8.1 Planned uses of the database

The database system that we designed is for medical research and further drug discovery. With the database, researchers can be easily access to the extremely large data sets stored in the database and extract any related particular information from he database. For example, there are four kinds of main data domains including genes, diseases, cell lines and perturbations. Users can know a gene's id, symbol and title by type in a gene name into the input field. In the same way, a user can also see a perturbation's type, a disease's description and a cell line's detail information. While all of this information is accessible separately, this database is target specifically at making connections in a meaningful, practical way which lends itself to speeding up experimental design and execution.

8.2 Potential areas for added functionality

There are still many functionalities available to be added into the current app. In the future, we can continuously amplify its function like adding more new data domains into the database, enlarge the scale of the existing data sets, and improve the running speed and system safety.

Currently, the way that our publication table is connected to the rest of the database is rudimentary– it is connected with the four main data domains, but there is not a way to easily add key words, more connected genes/cells/perturbations, or other supplemental materials such as figures, summary results, and so on. Seeing published results about perturbations is crucial for making a decision for which research targets to pursue.

Another idea we had was support users uploading their own data files into the database and generating a comparison result after comparing part of the file with the database in the information. For example, if the user uploads a list of new perturbations, the system may return a ranked list of most relevant genes and disease. For further use, the well-developed app should also have gene expression values recorded. Biologists use gene expression values to evaluate the reaction of the gene after receiving a perturbation. Along this line of thought, the app has potential to be used for recording expression value and using this value to predict a perturbation's effect on different disease or gene in various cell lines.

A series of visualization function would also be helpful to show data to the users in a more meaningful and readable way. Administrators can design several visualization modes for people to extract data with multiple choices of chart type.

9 ACKNOWLEDGEMENTS

Thank you to Dr. Kathleen Durant for your instruction!

Stock image for the background was provided free use from https://www.pexels.com/photo/close-up-of-microscope-256262/ d